DOCUMENT RESUME

ED 187 176 HE 012 564

AUTHOR Overall, J. U.; Marsh, Herbert W.
TITLE Relative Influence of Course Level, Course Type, and Instructor on Students' Evaluation of Instruction.
PUB DATE 80
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 1980)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty; *College Instruction; College Students; Comparative Analysis; Educational Research; Graduate Students; Higher Education; *Predictor Variables; *Student Evaluation of Teacher Performance; Teacher Behavior; *Teacher Characteristics; *Teacher Effectiveness; Undergraduate Students

ABSTRACT

The relative contribution of course level (undergraduate versus graduate), course type (accounting, economics, management theory, etc.) and the specific instructor on students' evaluation of instruction was investigated. College students (N=1374) evaluated teaching effectiveness both at the end of each course (N=100) and again one year after graduation. Mean end-of-term ratings were similar to those collected after graduation, and the two sets of ratings were highly correlated. The variance attributable to the specific instructor was much larger than that due either to course level or course type for both end-of-term and followup ratings, and the effects were stable. These findings are claimed to show that the particular course one teaches or the level at which it is taught is relatively less important in determining the outcome of student ratings than who teaches it. (Author/SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187176

RELATIVE INFLUENCE OF COURSE LEVEL, COURSE TYPE, AND INSTRUCTOR ON STUDENTS' EVALUATION OF INSTRUCTION

J.U. Overall

California State University, Dominguez Hills

Herbert W. Marsh

University of Southern California

College students (N=1374) evaluated teaching effectiveness both at the end of each course (N=100) and again one year after graduation. Mean end-of-term ratings were similar to those collected after graduation, and the two sets of ratings were highly correlated (median r = .83). The purpose of this study was to determine the relative contribution of course level (undergraduate vs graduate), course type (accounting, economics, management theory, etc.), and the specific instructor in determining evaluations of teaching. The variance attributable to the specific instructor was much larger than that due either to course level or course type for both end-of-term and follow-up ratings, and the effects, were stable. These findings show that the particular course one teaches or the level at which it is taught is relatively less important in determining the outcome of student ratings than who teaches it.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. U. Overall

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

HE 012 561

Relative Influence of Course Level, Course Type, and Instructor on Students' Evaluation of Instruction

Determination of the relative contribution of the course being evaluated and the specific instructor teaching it is important to the interpretation of student evaluations of teaching. In particular, if the effect of the course is large, relative to the effect of the instructor, then it would be difficult to justify the comparison of ratings obtained by different teachers teaching different courses.

Researchers have considered alternative approaches to the problem of the relative influence of the specific instructor and of the course he/she teaches (Bausell, Schwartz & Purohit, 1975; Gillmore, Kane & Naccarato, 1978; Kulik & Kulik, 1974). Conceptually, the problem can be approached from generalizability theory, which is based upon variance estimates within an analysis of variance framework (Kane & Brennan, 1977). However, the actual application of this approach is confounded by many technical difficulties. Ideally, the approach would require that each of a set of courses be taught by each of a set of instructors and that each would be evaluated by the same set of students (with no missing data). In fact, different courses are typically taught by different instructors and are evaluated by different sets of students. While, over a several-year period of time, it is likely that several different instructors will teach a given course, it is questionable whether the inherent teaching abilities of these different instructors represent a random sample. Similarly, it is unlikely that the students typically taking any given course represent a random sample of the population of student types. It is even more unlikely that the different courses that are taught by any particular instructor represent a random sample of course types.

In the present study all students were required to take all the courses offered in a given program. This eliminates the problem of self selection that causes systematic differences in the samples of students that enroll in different courses. Furthermore, since student evaluations were collected at both the end of each class and one year after graduation from the program, the course and teacher effects can be assessed separately for both sets of data, and the stability of the

effects over time can be determined.

Method

Subjects were undergraduate and graduate students in the School of Management at California State University, Dominguez Hills, who completed degree requirements between 1975 and 1977. During this period students from randomly selected classes were asked to put their student identification numbers on all evaluation forms completed during the last week of class. One year after graduation follow-up surveys were sent to these same students asking them to again evaluate the same teachers. The end-of-term survey contained 28 items, while the follow-up survey contained a subset of 9 of these items. Responses on both surveys were made along a 9 point response scale that varied from "1-very low/never" to "9-very high/always".

End-of-term surveys were completed by 83%; 65% returned the follow-up surveys; and 56% completed both. Analysis was performed on the ratings from 100 sections (out of a possible 124) in which at least 8 students completed both end-of-term and follow-up surveys. A total of 1374 successfully matched end-of-term and follow-up ratings were obtained. Of the potential 24,732 responses (9 end-of-term and 9 follow-up ratings for each set of ratings), there were only 31 missing values. Furthermore, preliminary analyses indicated that end-of-term ratings for those students completing both the end-of-term ratings and the follow-up ratings did not differ significantly from those of students completing only the end-of-term ratings.

There were ratings of 19 different courses; 11 graduate courses and 8 undergraduate courses. Each course was evaluated by at least three different sections of students. (An average of 5.05 sections evaluated each course). Different sections of the same course were taught by an average of three different instructors. A total of 40 different instructors were included in the study. Each of these instructors taught an average of 2.5 different sections, generally different sections of the same or similar courses. The 19 different courses were grouped into five content categories: Quantitative Analysis, Economics, Accounting, Management, and Other (Business

Communication, Social & Ethical Issues, Legal Aspects, etc.). Each category consisted of at least three different courses and contained courses at both the graduate and undergraduate levels.

Several different analyses of variance were performed to assess the effects of the program level (graduate and undergraduate), the course type, the different items, the specific instructor, and the stability of these effects over time. Since the effects of program level, course type, items, and time were crossed (e.g., all possible combinations occurred), all their interactions could be determined. The effect of the specific instructor was nested within both program level and course type; consequently, the main effect of the specific instructor also contains variance attributable to interactions between the specific instructor and these other two effects (but see Gillmore, et al., 1978). The interactions between the contribution of the specific instructor and the effects of items and time could be determined. The actual analysis was performed using the commercially available SAS programs (SAS, 1979, pp. 121-129).

Statistical significance, because of the nature of the study and the large sample size, is less important in this study than is the variance accounted for by each effect. Since variance components could not be computed for this study (see footnote 1), eta squared was used as an index of the strength of each effect (Kirk, 1968).

Results

The results showed both relative and absolute agreement between end-of-term and retrospective ratings (see Marsh & Overall, 1979; Overall & Marsh, in press; for further discussion). Based upon class-average responses, the median correlation (relative agreement) between end-of-term and retrospective ratings was .83. Mean differences (absolute agreement) between the two sets of ratings were small; the median rating across the nine items was 6.63 for retrospective ratings and 6.61 for end of term ratings. These findings demonstrate that students' evaluations are remarkably stable over time. Of particular interest in this study, however, is the

relative contribution of the specific instructor and the course in determining the ratings, and whether these effects are also stable over time. A series of analyses of variance were undertaken to answer these questions.

The first analysis of variance performed contained all five effects (Program Level, Item, Course Type, Specific Instructor, and Time of Evaluations) and all possible interaction effects (see Table I). Inspection of the ANOVA table indicates that while many of the effects were statistically significant, only the contribution of the Specific Instructor --which explained 15.5% of the variance -- accounted for more than 2.5% of the data. The effect of the program level was not statistically significant, and the course type accounted for only 1.5% of the data. The effects of program level and course type did interact, but accounted for only an additional 2.2% of the variance. The effect of a specific instructor varied somewhat with the time at which the evaluations were collected, but this interaction accounted for only 1% of of the variance. In fact, the time at which evaluations were collected -- either singly or in combination with other variables -- accounted for very little of the variance, a further indication of the stability of the ratings.

Two similar analyses of variance were performed separately on the end-of-term and on the follow-up ratings. These results (see Table II) also indicate that the teacher effect was consistently much larger than either the effect of the program level or the course type. Each of the effects -- particularly the teacher effect -- tended to be larger for the retrospective follow-up ratings.

In the final set of statistical computations, 18 separate analyses of variance were performed. An analysis of each item was performed separately for the end-of-term and the follow-up ratings. These analyses (see Table III), also indicate the same pattern of results; the effect of the specific instructor was consistently much larger than any other effect. This pattern was found in every individual item for both the end-of-term and follow-up evaluations, although the magnitude of the effects did vary somewhat with different items.

Discussion

Results of each of the different sets of analyses have demonstrated a consistent pattern of findings. The effect of program level on student ratings was not statistically significant. The effect of course type, while statistically significant, accounted for no more than 2-3% of the variance in any of the analyses. Furthermore, if differences due to individual instructors were considered to be a random effect and were used as an error term in assessing the significance of the course type effect, as suggested by Gillmore (1977), the effect of course type would not even be statistically significant.

In contrast to the small effects of both the program level and the course type, the effect of the specific instructor was large, generally accounting for 5 to 10 times as much variance as either course type or program level. A similar pattern of findings was found in both the end-of-term and follow-up ratings, although the specific instructor effect tended to be even larger in the follow-up ratings.

The results of this study have important implications for the use and interpretation of students' evaluations of teaching effectiveness. These evaluations were shown to be remarkably stable over time, offering further evidence of their validity. More specifically, however, the findings demonstrate that the particular course taught is relatively much less important in determining the outcome of the student ratings -- either at the end of the course or in retrospect -- than the specific instructor who teaches it.

FOOTNOTES

1--Because of the theoretical work done by Cronbach (Cronbach, Gleser, and Nada, 1972) and its application to student evaluatons (Kane & Brennan, 1977), the most useful measure of the strength of an effect would have been variance components and generalizability coefficients. However, their computation requires that the design is completely balanced, as well as having equal sample sizes. Neither of these conditions existed for the present study, and not even a commercially available statistical procedure with an iterative procedure for estimating variance components (SAS, 1979) was able to determine a solution.

References

Bausell, R.B., Schwartz, S., and Purohit, A. An examination of the degree to which various student parameters replicate across time. Journal of Educational Measurement, 1975, 12, 273-280.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The Dependability of Behavioral Measurements. New York: John Wiley & Sons, 1972.

Gillmore, G.M. How large is the course effect? A note on Romney's course effect vs. teacher on students' ratings of teacher competence. Research in Higher Education, 1977, 7, 187-189.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. The generalizability of student ratings of instruction: Estimating the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-15.

Kane, M.T. and Brennen, R. The generalizability of class means. Review of Educational Research, 1977, 47, 267-292.

Kirk, R.E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole, 1968.

Kulik, J.A. and Kulik, C.C. Student ratings of instruction. Teaching of Psychology, 1974, 1,51-57.

Marsh, H.W. and Overall, J.U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". Research in Higher Education, 1979, 10, 1139-147.

Overall, J.U. and Marsh, H.W. Students' evaluations of instruction: Stability at both the undergraduate and graduate levels. Journal of Educational Psychology, (In Press).

SAS Institute, Statistical Analysis System User's Guide (1979 Edition). Raleigh, N. C. : SAS Institute, 1979.

Smith, P.L. The generalizability of student ratings of courses: Asking the right questions. Journal of Educational Measurement, 1979, 16, 77-87.

Smith, P.L. The stability of teacher performance in the same course over time. Research in Higher Education, 1979, 11, 153-165.

FOOTNOTES

1--Because of the theoretical work done by Cronbach (Cronbach, Gleser, and Nada, 1972) and its application to student evaluatons (Kane & Brennan, 1977), the most useful measure of the strength of an effect would have been variance components and generalizability coefficients. However, their computation requires that the design is completely balanced, as well as having equal sample sizes. Neither of these conditions existed for the present study, and not even a commercially available statistical procedure with an iterative procedure for estimating variance components (SAS, 1979) was able to determine a solution.

References

Bausell, R.B., Schwartz, S., and Purohit, A. An examination of the degree to which various student parameters replicate across time. Journal of Educational Measurement, 1975, 12, 273-280.

Cronbach, L.J., Gleser, G.C., Nanda, H., and Rajaratnam, N. The Dependability of Behavioral Measurements. New York: John Wiley & Sons, 1972.

Gillmore, G.M. How large is the course effect? A note on Romney's course effect vs. teacher on students' ratings of teacher competence. Research in Higher Education, 1977, 7, 187-189.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. The generalizability of student ratings of instruction: Estimating the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-15.

Kane, M.T. and Brennen, R. The generalizability of class means. Review of Educational Research, 1977, 47, 267-292.

Kirk, R.E. Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole, 1968.

Kulik, J.A. and Kulik, C.C. Student ratings of instruction. Teaching of Psychology, 1974, 1,51-57.

Marsh, H.W. and Overall, J.U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". Research in Higher Education, 1979, 10, 1139-147.

Overall, J.U. and Marsh, H.W. Students' evaluations of instruction: Stability at both the undergraduate and graduate levels. Journal of Educational Psychology, (In Press).

SAS Institute, Statistical Analysis System User's Guide (1979 Edition). Raleigh, N. C. : SAS Institute, 1979.

Smith, P.L. The generalizability of student ratings of courses: Asking the right questions. Journal of Educational Measurement, 1979, 16, 77-87.

Smith, P.L. The stability of teacher performance in the same course over time. Research in Higher Education, 1979, 11, 153-165.

TABLE I

Analysis of Variance Table For All Effects Combined : Effects of Program Level (Undergraduate or Graduate), Course Type, Specific Instructor, Different Evaluation Items, and Time (End-of-term or Retrospective Follow Up ratings)

Source	DF	ANOVA SS	F-Ratio	ETA SQ x 100%
Program Level	1	1.8	0.57	0.0%
Item	8	212.8	8.66**	0.2%
Time	1	18.5	6.03*	0.0%
Course Type	4	1463.9	119.19**	1.5%
Prog X Item	8	22.4	0.91	0.0%
Prog x Time	1	1.4	0.45	0.0%
Prog x Crse Type	4	2097.9	170.8**	2.2%
Time x Item	8	147.7	6.01**	0.2%
Item x Crse Type	32	330.8	3.37**	0.3%
Time x Crse Type	4	133.9	10.90**	0.1%
Prog x Time x Item	8	46.4	1.89	0.0%
Prog x Item x Crse Type	32	339.9	3.46**	0.4%
Time x Item x Crse Type	32	41.3	0.46	0.0%
Prog x Time x Item x Crse	36	40.9	0.37	0.0%
Instructor	39	14894.6	124.38**	15.5%
Instructor x Item	312	1842.0	1.92**	1.9%
Instructor X Time	43	974.9	7.38**	1.0%
Instructor x Time x Item	308	470.3	0.50	0.5%
Total Model (combined influ- of all the above effects)	881	23081.49	28.53	24.0%
Error	24701	73141.57	----	76.0%
TOTAL (sum of Total Model and Error)	25582	96223.06	----	100.0%

* $p < .05$, ** $p < .01$

NOTE: A five-factor analysis of variance was used to assess each effect, and the combined effect (Total Model). In this design, the specific instructor was nested within the Program level and the Course type. Consequently, higher order interaction effects with these two variables and the specific instructor effect were confounded.

TABLE II

Analysis of Variance Table For All Effects Combined :
Effects of Program Level (Undergraduate or Graduate),
Course Type, Specific Instructor, Different Evaluation Items,
and Time (End-of-term or Retrospective Follow Up ratings)

Source	DF	End-of Term Ratings ANOVA SS	F-Ratio	ETA SQ x 100%	Follow-up Ratings ANOVA SS	F-Ratio	ETA SQ x 100%
Program Level	1	0.1	0.11	0.0%	3.1	1.08	0.0%
Item	8	265.9	10.31**	0.5%	94.6	4.05	0.2%
Course Type	4	953.9	74.00**	1.9%	643.8	55.15**	1.3%
Prog X Item	8	37.9	1.47	0.0%	30.9	11.32	0.0%
Prog x Crse Type	4	944.9	73.3**	1.9%	1160.3	999.39**	2.4%
Item x Crs Type	32	242.8	2.35**	0.5%	129.3	1.39	0.3%
Prog x Item x Crse Type	32	209.3	2.03**	0.4%	164.3	1.76**	0.3%
Instructor	39	5994.3	47.7**	12.4%	9867.9	86.70*	20.7%
Instructor x Item	312	1418.1	1.41**	2.9%	901.5	0.99	1.9%
Total Model (Combined influence of all the above effects)	440	10067.2	7.10	20.8%	12995.8	10.12**	27.2%
Error	11909	38380.1	----	79.2%	34761.5	----	72.8%
TOTAL (sum of Total Model and Error)	12349	49471.3	-----	100.0%	477573.3	----	100.0%

* $p < .05$, ** $p < .01$

NOTE: Two four-factor analyses of variance were used to assess each effect, and the combined effect (Total Model). In this design, the effect of the specific instructor was nested within the Program level and the Course type. Consequently, higher order interaction effects with these two variables were confounded.

APPENDIX I

Long Term Stability of Students' Evaluations: Relative and Absolute Agreement Between End-of-term & Follow-up Ratings

Evaluation Items	RELATIVE AGREEMENT: Correlations Between End-Of-Term & Follow-Up Individual Student Responses	Class Average Responses	ABSOLUTE AGREEMENT: Mean Differences Between End-Of-Term & Follow-Up End of Term	Follow Up	Difference
Purpose of class assignments made clear	.55	.81	6.66	6.71	+0.05
Course objectives outlines	.58	.84	6.56	6.69	+0.13**
Class presentations prepared & organized	.62	.79	6.66	6.76	+0.10*
You learned something of value	.53	.81	6.94	6.75	-0.19**
Instructor considerate of your point of view	.58	.83	6.94	6.65	-0.29**
Instructor involved you in class discussions	.56	.84	6.81	6.68	-0.13**
Instructor stimulated your subject interest	.58	.82	6.57	6.45	-0.12*
Overall Instructor Rating	.65	.84	6.80	6.63	-0.17**
Overall Course Rating	.56	.83	6.57	6.73	+0.16**
MEDIAN ACROSS ALL 9 RATING ITEMS	.58	.83	6.66	6.69	

* $p < .05$, ** $p < .01$

Note: A total of 1374 student responses from a total of 100 different sections each rated instructional effectiveness at the end of each class (End Of Term) and again one year after graduation (Follow Up).

TABLE III

Variance Accounted For By Program Level (Undergraduate or Graduate), Course Type, and Specific Instructor: Separate Analyses Were Performed For End-of-term (EOT) and Follow-up (FU) Ratings of Each Item

EVALUATION ITEMS		Program Level F(1,1325) = (% Variance)	Course Type F(4,1325) = (% Variance)	Prg X Crse Interaction F(4,1325) = (% Variance)	Instructor F(39,1325) = (% Variance)	Total Model Combined Effects F(48,1325) = (%Variance)
Purpose of class assignments made clear	EOT	0.1 (0.0%)	8.3 (2.0%) **	8.0 (1.9%) **	5.9 (14.1%) **	6.1 (18.2%) **
	FU	0.1 (0.0%)	8.7 (1.9%) **	9.7 (2.1%) **	11.5 (24.3%) **	10.9 (28.3%) **
Course objectives outlined	EOT	0.1 (0.0%)	8.8 (2.1%) **	10.4 (2.4%) **	7.8 (17.9%) **	8.0 (22.4%) **
	FU	1.5 (0.1%)	8.2 (1.8%) **	12.7 (2.7%) **	11.4 (24.0%) **	11.0 (28.5%) **
Class presentations prepared & organized	EOT	3.2 (0.2%)	13.7 (3.1%) **	10.1 (2.3%) **	8.8 (19.4%) **	9.2 (25.0%) **
	FU	0.0 (0.0%)	8.2 (1.7%) **	15.9 (3.3%) **	12.0 (24.8%) **	11.8 (29.9%) **
You learned something of value	EOT	0.1 (0.0%)	5.5 (1.4%) **	7.7 (1.9%) **	4.8 (11.9%) **	5.0 (15.3%) **
	FU	0.2 (0.0%)	4.1 (0.9%) **	12.2 (2.8%) **	9.3 (20.7%) **	8.9 (24.5%) **
Considerate of your point of view	EOT	2.9 (0.2%)	10.7 (2.5%) **	11.7 (2.8%) **	6.7 (15.7%) **	7.4 (21.2%) **
	FU	1.2 (0.1%)	7.8 (1.8%) **	12.7 (2.9%) **	9.1 (20.1%) **	9.1 (24.8%) **
Instructor involved you in class discussion	EOT	2.0 (0.1%)	19.5 (4.5%) **	12.1 (2.8%) **	7.0 (15.9%) **	8.3 (23.4%) **
	FU	3.6 (0.2%)	10.3 (2.3%) **	10.7 (2.4%) **	9.3 (20.5%) **	9.4 (25.4%) **
Instructor stimulated your subject interest	EOT	0.8 (0.1%)	10.4 (2.6%) **	8.5 (2.1%) **	5.6 (13.4%) **	6.1 (18.1%) **
	FU	0.0 (0.0%)	6.3 (1.4%) **	10.2 (2.3%) **	10.5 (22.7%) **	9.9 (26.4%) **
Overall Instructor Rating	EOT	0.6 (0.0%)	11.0 (2.7%) **	13.0 (3.2%) **	5.8 (13.7%) **	6.7 (19.6%) **
	FU	4.4 (0.2%) *	6.5 (1.3%) **	17.5 (3.6%) **	13.1 (26.3%) **	12.7 (31.5%) **
Overall Course Rating	EOT	2.4 (0.1%)	3.9 (1.0%) **	8.1 (2.0% **	6.5 (15.6%) **	6.4 (18.7%) **
	FU	0.6 (0.0%)	7.2 (1.7%) **	10.9 (2.5%) **	8.6 (19.3%) **	8.5 (23.4%) **
MEDIAN ACROSS ALL 9 RATING ITEMS	EOT	0.1%	2.5%	2.3%	15.6%	19.6%
	FU	0.0%	1.8%	2.7%	22.7%	26.4%

* p<.05, **p<.01

Note: A three factor analysis of variance was used to assess each effect and the combined effect (total model). Variance estimates were determined by taking a ratio of the SSeffect/SStotal (Eta squared).